DOCUMENT RESUME

ED 063 333 TM 001 351

AUTHOR Kriewall, Thomas E.
TITLE Aspects and Applications of Criterion-Referenced Tests.
INSTITUTION Institute for Educational Research, Downers Grove, Ill.
REPORT NO TP-103
PUB DATE Apr 72
NOTE 27p.; Paper presented at the annual meeting of the American Educational Research Association (Chicago, Illinois, April 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Ability Grouping; Academic Performance; *Criterion Referenced Tests; Decision Making; *Evaluation Methods; Individualized Instruction; *Instructional Design; *Item Sampling; Measurement Techniques; Models; *Student Grouping; Test Reliability; Test Validity

ABSTRACT

The measurement information generated by CRT's is designed for use in instructional management systems where classifications of pupils for treatment are to be decided on the basis of minimal data consistent with predetermined limits for the errors of misclassification. The measures obtained are content specific estimates of proficiency useful for the stratification of learning groups on a day-to-day basis if need be. By sampling across items rather than across persons, absolute measures of proficiency are obtained which can be reliably interpreted for nonrandomly selected pupils, the pupils of particular concern. The model is designed for wide variety of applications but retains in the concept of proficiency a simple and useful index for instructional management. The empirical data generated have clear implications for instructional decision-making. Some of the applications of item-sampling theory include the ability to (1) categorize learners into temporary learning groups on the basis of a common requirement for instructional treatment; (2) assess the relative effectiveness of competing instructional treatments; (3) determine, in the case of established instructional segments having predetermined performance standards, which individuals require further prescriptive assistance; and (4) in the case of curriculum development, to indicate hierarchical relations within a content sequence. (Author/DB)

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
INSTITUTE FOR EDUCATIONAL RESEARCH
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER.

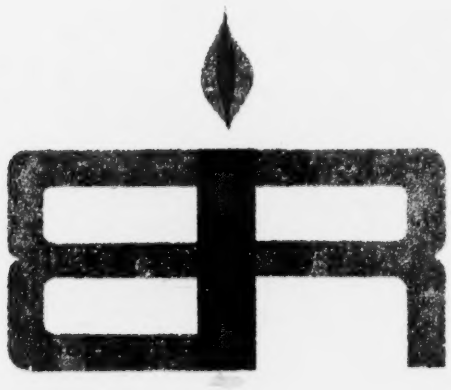
ED 063333

ASPECTS AND APPLICATIONS OF CRITERION-REFERENCED TESTS

Technical Paper #103

Thomas E. Kriewall, Ph.D.
Research Analyst

A paper presented at the annual meeting of the
American Educational Research Association
April 3-7, 1972 Chicago, Illinois



INSTITUTE FOR EDUCATIONAL RESEARCH
1400 West Maple Avenue
Downers Grove, Illinois 60515

TM 001 351

ABSTRACT

In summary, the measurement information generated by CRT's is designed for use in instructional management systems where classifications of pupils for treatment are to be decided on the basis of minimal data consistent with predetermined limits for the errors of misclassification. The measures obtained are content-specific estimates of proficiency useful for the stratification of learning groups on a day-to-day basis if need be. By sampling across items rather than across persons, absolute measures of proficiency are obtained which can be reliably interpreted for nonrandomly selected pupils, the pupils of particular instructional concern. The model is designed for wide variety of applications but retains in the concept of proficiency a simple and useful index for instructional management. The empirical data generated have clear implications for instructional decision-making.

Some of the applications of item-sampling theory include the ability to (1) categorize learners into temporary learning groups on the basis of a common requirement for instructional treatment (Diagnosis and Prescription Function); (2) assess the relative effectiveness of competing instructional treatments (Instructional Assessment Function); (3) to determine, in the case of established instructional segments having predetermined performance standards, which individuals have acquired minimal standards of proficiency required for mastery and which learners require further prescriptive assistance (Quality Control Function); and (4) in the case of curriculum development, to indicate hierarchical relations within a content sequence (Curriculum Design Function).

© INSTITUTE FOR EDUCATIONAL RESEARCH
March, 1972

ASPECTS AND APPLICATIONS OF CRITERION-REFERENCED TESTS

I. THE INSTRUCTIONAL CONTEXT FOR CRITERION TESTING

Criterion-referenced testing might be considered to be in its adolescence, confused but promising. The fact that three theories are being presented at this symposium is belated evidence that CRT implementation, indeed testploitation, has far outdistanced conceptualization. In today's profusion of claims we are driven back to consider fundamentals of definition and properties that mark the criterion-referenced test as a distinct breed with a unique purpose in education.

I propose to set a discussion and definition of criterion tests in the context of classroom needs that have created much of the interest in the theory at this time. The primary source of interest, in my view, is related to the growing implementation of individualized curricula.

The trend toward individualization of instruction has forced changes in many educational practices. Traditional testing and grading practices, however, have not been readily adapted to many instructional innovations adopted in recent years.

One of the evaluation problems faced by those concerned with individualization of instruction is that the classical norm-referenced test (NRT) is built, to use MacDonald's (1965) term, on a "mythology" that is inapplicable or irrelevant to many new instructional problems. In explanation

of the term "mythology," MacDonald says in part:

> . . . we may utilize many metaphors in our talk. . . Some of these metaphors have been raised to the level of myths. They are myths by definition here because they are used to prescribe... when in reality they are only possible ways of viewing, with uncertain probabilities of validity.

In much the same sense it is possible that new metaphors are needed to clarify some evaluation problems which, as Glaser (1963) has indicated, have been clouded over by an entrenched NRT mythology.

Classical test metaphors have arisen as rationales or interpretations for procedures and assumptions initially adopted mainly on theoretical grounds. For example, $\pi_g = \mathcal{E}_a\ (Y_{ga})$ is a well-defined theoretical construct. Such a defining statement has been called a syntactic definition (Lord and Novick, 1968, p.15). An empirical, behavioral, or semantic meaning such as "item difficulty" is what Carnap (1950) has called the explication of the construct. The term metaphor is used here because a change of context can render a given explication irrelevant. Metaphors, raised to a level at which they become an unchallenged basis for prescribing test construction procedures when in fact other alternatives may be just as or even more useful, are myths. General and uncritical acceptance of myths leads to faulty test construction and confusion. What is needed is a relevant set of metaphors. Item-sampling theory provides such a set for criterion test construction.

Some problems in applying norm-referenced metaphors to criterion test construction include the following.

Item Difficulty

Item difficulty is defined as the expected relative score on an item by a population of examinees. It is often denoted by the symbol p (or pi)

because of its interpretation as a probability. If an individual is selected at random from the population of examinees, then p is the probability such a person will respond correctly to the item. A difficulty with item difficulty, from the teacher's point of view, is that at the local level one is not teaching a random sample of children selected from a specified population, but rather a particular group of individuals. Inferences must be made concerning the performance capabilities of these nonrandomly selected individuals. It is essential in the context of day-to-day instruction that particular individuals be treated as such and not as a random sample from some larger population.

This suggests that item difficulty is not an appropriate or particularly useful concept in its classical sense. A new metaphor is required. The requirements of the instructional problem suggest what this metaphor should be.

The goals of instruction can frequently be cast in terms of developing specified levels of performance on certain categories of tasks. At any point in time, the teacher may be trying to develop a delimited set of performance behaviors. A pupil may develop the desired behavior in various ways. He may completely fail to comprehend the ideas involved. Or he may develop specialized techniques which work on some problems but not all of a given class. Or he may learn general procedures that render all problems of a given class equally capable of solution, subject only to random human failures due to personal or environmental sources of error. What the teacher needs to know at given points in time is the probability for success that a given pupil has with respect to a specified class of performance tasks.

Rather than sample performance across a hypothetical population of pupils, it is more appropriate to measure the individual's behavior on a random sample of problems drawn from a clearly defined population of tasks. The individual's relative score (i.e. percentage score) can then be interpreted as an estimate of his proficiency relative to the defined item population. Proficiency is a metaphor of essential importance in criterion test theory. It restores percentage scores to a place of prominence at the cost of changing item difficulty from a definition to an assumption. The assumption is that an individual's proficiency is constant at a given point in time for all tasks of a given class. In other words, all the problems in the class are of equal "difficulty" for the individual.

I (pragmatically) define tests constructed to provide proficiency measures, as described above, criterion-referenced tests or CRT's.

Content Validity

It has been said that the mental traits a test measures is a question which the psychometrician has no adequate way of answering (Lord and Novick, 1968, p.528). Because of this haziness, classical item selection procedures serve only as a guide rather than an algorithm for test construction. In the final analysis, the test builder must make subjective decisions concerning a given item's relation to whatever it is he wants to measure.

However, the usefulness of a criterion-test is vitiated unless the test has obvious content validity (Ebel, 1962). It is of little use to an instructional manager to know a pupil is 90% proficient, for example, if it is not known what specific content or skills compose the proficiency. The item-sampling model described here, therefore, begins with the assumption of prima facie content validity. The essential metaphor that enables

one to meet this condition is the notion that a learning objective (LO) is defined by a specified item population.

If a test contains a collection of items on a wide variety of topics, one can conclude from the test results very little except the degree to which the students are capable of retaining crammed-in bits of unrelated knowledge. However, if the test is constructed by random sampling from a specific class of problems (e.g. situations requiring grammatical analysis, computation, or other disciplined patterns of reasoning), then it follows that the instruction should be directed to the development of relevant skills in sufficient generality that a uniformly high probability exists for successful behavior no matter which particular problem is selected for test purposes.

The dependence of classical test construction on subjective decisions made by the test builder is undesirable in criterion-test construction not only because of its deleterious effects on content validity but also because of a need that exists in individualized instructional systems for the generation of many "parallel" tests (Hively et al, 1968). The mastery paradigm for instruction entails the possibility of an individual repeatedly recycling through a given body of content (usually over an extended period of time as other learning goals are interleaved). Upon the completion of each cycle, no matter how the sequencing problem is handled, one needs a new version of the test to determine if the pupil has finally achieved some minimal level of mastery.

One technique for generating many parallel versions of a given criterion-test is the use of the random number generator of a computer to sample the specified item population. In particular content areas, such

as mathematics, it is further possible to generate the items rather than to recall them randomly from a prepared list. However, in most subject areas, one must resort to random selection from an existing item pool. (Human error in item preparation can contribute substantially to CRT measurement error in the latter case. In my judgment, collections of items related to frequently used learning objectives are not of sufficient size or quality to support widespread use of CRT measurement at this time. Major efforts are therefore required in this area before the state-of-the-art can be much improved.)

The Assumption of Normality

Another metaphor, commonplace in classical test construction, is that tests measure one or more mental traits and that these traits are normally distributed among a population of examinees. However, the assumption of a normal distribution for proficiency is clearly contradictory to the purpose of instruction. It negates the prospect of mastery in fundamentals.

It seems more likely that proficiency distributions following instruction are multimodal, and probably essentially bimodal or trimodal. Normal distributions occur about the modal points only because of random factors generally classes as "error." The data of interest to the teacher are not the class mean and relative ranking of class members but rather data which permits the correct classification of students into subsequent instructional groups, together with estimates of absolute levels of proficiency within each group. These data would be sufficient to assist the teacher in making instructional decisions in regard to differentiating instruction and in comparing the effectiveness of alternative instructional treatments.

Test Reliability

Reliability is defined in classical test theory as the squared correlation between true and observed scores. The metaphor used to give meaning to this definition is that reliability measures the extent to which a repeated measure would agree with the original measure on a group of examinees. If the examinees' traits measured by the test have not changed, then another administration of a test which measures the same traits should ideally provide the same score for each examinee.

The metaphor becomes a myth when it is used to prescribe methods of test construction. It is easy to show that maximum variance is achieved when item difficulties are approximately 0.50. Thus, for maximum test reliability, it is commonly recommended that use of items with either very low or very high p-values be avoided.

The problem with this procedure for CRT design has been already indicated. The "difficulty" of items for a nonrandomly selected group of persons is, first of all, not known before the test is administered. In fact, it might be considered that the purpose of a CRT is to estimate the overall "difficulty" of the class of items for an individual.

II. FUNDAMENTALS OF A CRT THEORY

From these considerations one can derive the essentials of item sampling theory for CRT construction as follows:

Definition 1: A learning objective (LO) is a rule for generating a class of performance tasks, or alternatively a list of all performance tasks which comprise the objective.

Definition 1 asserts that a relevant situation for the use of a criterion-test is one in which it is possible to define, a priori, a population of performance tasks comprising a learning objective. For example, it may

be desired to test a pupil's proficiency in detecting whether or not a pair of randomly selected three-letter (nonsense) words are the same, where each pair of words is built on the pattern "consonant-vowel-consonant." One might further restrict the first and last consonant to have certain properties such as being the same within a word but randomly different or the same between pairs of words being tested. The "replacement set" from which the consonants and vowels are to be randomly selected can be specified as desired. In this way the set of tasks to be tested becomes well-defined and, by random sampling from the population, one can (1) estimate an individual's proficiency relative to the defined task population or (2) on the basis of minimum test size and specified limits of classification errors, classify individuals into groups which (a) have proficiency greater or equal to some minimal mastery criterion or (b) have proficiency less than or equal to some maximum nonmastery criterion.

Assumption 1: Each pupil has a single proficiency at any given point in time relative to a specified learning objective.

Assumption 2: Proficiency is a function of time.

Criterion tests, to be most useful, require application of a strict item sampling model. The term "strict" simply means that one first defines the item population, then selects a random sample of n items for test. This point is emphasized because it is at variance with conventional item-sampling techniques. Cornfield and Tukey (1956) have characterized the more usual approach as one involving first the choice of a sample on which statistical analyses are made then introducing an unspecified population of items "like those observed" for which inferences are to be made. With the same perspective, Lord and Novick (1968, p.234) speak in terms

of "n test items considered as a random sample from a population of items," rather than n items which are a random sample.

It is easy to see how, if the item pool is not delimited and defined a priori, the scaling falls to an interval level (at best) and why, in that case, one would have to resort to classical item selection procedures for building a measurement scale. With an unrestricted item-population, such as one consisting of items "like those" in a given sample, there is no evident limit to keep the item writer within the bounds of the learning objective. Given a relatively free hand in the exercise of his art, tests built with unrestricted item populations would be indistinguishable in structure from norm-referenced tests. A zero true score could not be assumed to exist under such assumptions. Observed scores would be a function of the mean item-difficulty in the selected sample. By biasing the item selection process to favor items of a given difficulty, it would be possible in such cases to build tests having some predetermined class mean. Thus the absolute value of the observed score would not be meaningful. Only the ranks, and possibly the differences between ranks, would preserve their meaning when the item population is not well-defined.

By contrast, pupils not familiar with the problem solving skills involved in the items found on a given CRT will show a true zero proficiency. Since the item pool is well defined, one cannot search about endlessly in an infinite pool in search of items which discriminate at arbitrarily low ability levels. The corresponding argument holds at the high proficiency end.

Proficiency and True Score

In order to establish the syntactic definition of "proficiency," imagine the curriculum to be structured in terms of some network of LO's. Consider a generic element of the structure, LO_k. Now suppose that student a has completed some phase of work with respect to LO_k. Further imagine that we require the student to respond to all the items in the population of items defined by LO_k. The proportion of items to which the student exhibits a correct response is a measure of his proficiency.

Definition 2: The proficiency of the a^{th} student with respect to the k^{th} LO, denoted by the symbol ζ_{ak} is defined to be the relative true score of a on all n_k items.

In the statistical sense of the term, ζ is a parameter or "population value." It is also, by virtue of being a parameter, a constant value for a given item-population and individual, at a given point in time. Pragmatically, therefore, proficiency may be taken to mean the fixed probability of a correct response to an item randomly selected from the k^{th} LO's population of problems. In the same sense that the Π-value serves as a measure of a given item's difficulty for a pupil-population, ζ_a may be regarded as the mean difficulty of an item population for a given individual.

The complement of proficiency is termed the "error rate."

Definition 3: $\zeta'_{ak} = 1 - \zeta_{ak}$ (error rate)

A CRT Performance Model

The definitions formulated so far suggest a number of properties that may be elaborated now. The definition of the proficiency parameter, ζ, suggests that we think of the pupil responding to each item in the CRT sample with a fixed probability of correct response. This is consistent with our usage of proficiency in the singular form. It means, from another point of

view, that we assume that no "learning" occurs during the time of the test administration which affects the pupil's proficiency.

We also think of the student's responses being independent, that is, we assume that the outcome of any trial is independent of the outcome of every other trial on a CRT. This, of course, amounts to a restriction in the way we generate CRT items, e.g., we must not pyramid problems so that one particular problem holds the key to solving one or more other problems.

Formally, this assumption of independent responses is syntactically defined as follows:

Let n = number of items on the test.

U_{ga} = random variable denoting the a^{th} student's response to the g^{th} item (U_{ga} = 0 or 1).

u_{ga} = an observed value of U_{ga}

Then the "local independence" assumption is equivalent to imposing the condition:

(1) $$\text{Prob}\left(U_{1*} = u_{1*}, U_{2*} = u_{2*}, \ldots, U_{n} = u_{n*} \mid \zeta\right) = \prod_{g=1}^{n} \text{Prob}\left(U_{g*} = u_{g*} \mid \zeta\right)$$

The pragmatic implications of local independence extend beyond item writing or selection techniques. For a homogeneous proficiency group (i.e., one in which all members have ideally the same proficiency, ζ), local independence says that erroneous responses occur randomly. Therefore, the item inter-correlations calculated using response data gathered from such a group will have an expected value of zero. This property can be used to identify homogeneous learning groups in a heterogeneous class using KR-20 as an index of homogeneity. When instructional grouping practices are appropriate, the groups may be formed by iterative procedures which begin with the highest or lowest scores, successively adding adjacent score groups until

the true variance estimated by KR-20 possess some suitable cutting point. A value of KR-20 = 0.10 seems to work well in this application (Kriewall, 1969)

To summarize the model, a student's CRT performance may be viewed as a sequence of independent Bernoulli trials, each having the same probability of success, ζ_{ak}.

Hence it follows that if an individual were to be repeatedly given tests consisting of random samples of size n drawn from LO_k, his score distribution would be given by

$$(2) \qquad f(x_a) = \binom{n}{x_a} \zeta_a^{x_a} (1-\zeta_a)^{n-x_a}$$

where the scoring formula is given by

$$(3) \qquad x_a = \sum_{g=1}^{n} u_{ga} \qquad \text{(raw score = sum of item scores weighted 0,1)}$$

and

(4) $f(x_a)$ = relative frequency of occurrence of test score x_a

$$(5) \qquad \binom{n}{x_a} = \frac{n!}{x_a!\ (n-x_a)!} \quad , \text{ the binomial coefficient}$$

According to this Bernoulli model, each examinee responds to an item as though he were tossing a coin with bias ζ_a.

CRT Statistics

The following are well-known properties of tests built according to the item-sampling model:

1. The observed test score, x_a, is a sufficient statistic for estimating ζ_a. "Sufficient" means that no information is lost by reducing the data given in the item-response vector

$$(6) \qquad \underset{\sim}{v} = (u_1, u_2, \ldots, u_n)$$

through use of the scoring formula given above in (3). Furthermore, if the

items are, in fact, parallel then x_a is the minimal sufficient statistic for estimating ζ.

2. Error of measurement is defined by

(7) $\eta_a = x_a - \underline{n}\ \zeta_a$

Since the expected value, $\mathcal{E}(x_a) = n\ \zeta_a$, it follows that the expected error over repeated testing of a given examinee at a given point in time is zero. Pragmatically, this means that a longer test of, say, m·n items, considered as a battery of m parallel tests each of length n, will provide a better true score estimate than a test of only n items provided that one does not make the test so long as to encounter error due to fatigue.

3. Error variance, for individual #a is determined by test length and proficiency:

(8) $\sigma^2(n_a) = n\ \zeta_a\ (1-\zeta_a)$

An estimate of error variance, unbiased over repeated item sampling, is derived from the relation

$$\hat{\sigma}^2 = \left(\frac{n}{n-1}\right)\ \sigma^2$$

$$\therefore\ \hat{\sigma}^2 = \left(\frac{n}{n-1}\right) \cdot n \left(\frac{x_a}{n}\right) \left(1 - \frac{x_a}{n}\right)$$

(9) or $$\hat{\sigma}^2(n_a) = \frac{x_a\ (n - x_a)}{n-1}$$

The variance of the error of measurement is a maximum when

$$\frac{\partial\ (\sigma^2)}{\partial \zeta} = n - 2n\zeta = 0$$

or when

$$\zeta = \tfrac{1}{2}$$

Criterion tests will therefore have least error of measurement when proficiency is near either extreme; error is largest when a student's

proficiency is in the middle range and presumably changing rapidly.

In order to reduce error due to bias, items should not be multiple choice since the expected score when $\zeta = 0$ is not $n\zeta$ but is determined by the number of distractors used. Constructed response items are highly desirable even though they are the source of economic problems in test construction at present.

III. IMPLICATIONS OF THE CRT MODEL

Item Selection

The item-sampling model described here as the paradigm for CRT construction is one of the simplest of test models. It places no condition on the items except, to preserve score meaning, all items must share at minimum the objective attributes which serve to characterize an LO. Mixing of items from different LO's results in the kind of confounding that would occur in any measurement if measures of different kind were combined into a single count.

The LO not only preserves score measuring, but also defines the scale of measurement. An absolute zero is the least possible proficiency; 1.0 is the maximum.

A measure of a CRT's "reliability" is the standard error of measurement, i.e., the standard deviation of the random sampling distribution of sample means given by

$$\text{S.E.M.} = \sqrt{\frac{\zeta\,(1 - \zeta)}{n}} \tag{10}$$

This is a measure of the accuracy with which the student's true proficiency is estimated by the test.

In general, CRT statistics are independent of item parameters and dependent only on test length, n, and ζ, the proficiency. By contrast, NRT reliabilities are functions of item parameters. For example, a lower limit to NRT reliability is given by coefficient alpha:

$$(11) \qquad \alpha = \frac{n}{n-1} \left(1 - \frac{\sum_{g=1}^{n} p_g q_g}{\sum_{g=1}^{n} \sigma_g \rho_{gX}} \right)$$

where n = number of items

p_g = item difficulty estimate

$q_g = 1-p_g$

ρ_{gX} = item-test correlation

σ_g = item variance

The basic reason why classical reliability formulas are complex functions of item parameters and CRT reliability is a simple function of pupil proficiency and test length lies in the distinct notions of reliability involved. The NRT needs not only reliable estimates of scores, but also maximum dispersions <u>between</u> scores in order to achieve replicable rankings. Differences in true scores must be magnified, so to speak, so that errors of measurement will not cause many inversions in rank to occur in test replications. The underlying assumption in the NRT is that one <u>expects</u> to find true differences in rank in any sample of persons. The psychometrician operates on the primary assumption that such differences are due to normally distributed differences in underlying traits, mental traits whose measure is his chief concern.

The CRT, as an instructor's tool, reflects the view of Carroll (1963) and Bloom (1968) that differences in native ability can be compensated for by individualizing the pace and method of instruction. Thus it is conceivable that among any given sample of persons, differences in proficiency may disappear as a result of instruction. Uniform achievement is the ideal expected. Thus one can be content to obtain replicable estimates of proficiency even though such a test could obviously fail to meet the additional requirement of ranking reliability which is characteristic of the NRT.

Perhaps most important to note is that, for the instructor, the instructional goals described by certain LO's are the <u>given</u> quantities, and the mental skills which account for proficiency are the learning variables. In other words, items are fixed and mental abilities are to be developed by instruction. In the NRT case, mental abilities of interest are the given aspects of reality; the task is to find items (the variables) which involve the use of an existing ability and thus measure the degree of its presence or absence. The latter is complicated by the fact that little is known of item/trait relationships. Thus one must rely on complicated inferences drawn from item data obtained in pilot testing with representative pupil groups.

<u>Minimal Test Length</u>

Consideration of instructional paradigms suggests the possibility of a conflict arising between the value systems of the instructor and the evaluator of instructional effectiveness (even if both functions are performed by one and the same person). An instructional model which is common to many current individualized instructional systems involves pre-test, instruction, and post-test. If, as is the usual case, a fixed

amount of time is available for the combined functions of instruction and evaluation, then the allocation of more time to one function necessarily decreases the time allotted for the other and a conflict exists. The instructor presses for more time in the hope of achieving higher levels of learning while the evaluator requires more time to either sample a greater range of objectives or to get better estimates of proficiency on a given selection of objectives (e.g., Walbesser and Carter, 1969).

The problem is to find an admissable, if not an optimum, solution to the conflict. The item sampling model is designed to be useful in the solution of this problem in two important ways.

The two viable options open are, first, to reduce test length while preserving efficiency* and, secondly, to use convergent testing strategies.+ The former can be handled through the test model in an analytical fashion; the latter by competent content analysis. Although the content analysis is in large part a judgmental matter, the test model helps focus attention on the central concerns by virtue of its emphasis on specified learning objectives. Consideration of the problem of minimizing test length leads to the use of acceptance sampling theory and methods of curtailing tests, such as Wald's Sequential Probability Ratio Test (SPRT) (Wald, 1947).

---

* "efficiency" is taken to mean "having adequately small probabilities for all relevant kinds of errors." (Birnbaum in Lord and Novick, 1968, p.436).
+ A convergent strategy depends on the existence of an inclusion relation between the ability being assessed and component abilities also of interest. For example, long division requires subtraction and multiplication operations to be performed in succession. Therefore, the measure of success on a test of division proficiency is a lower bound to the separate component proficiencies of multiplication and subtraction. Thus if long division proficiency is high, one can infer that both multiplication and subtraction proficiency are at least as high. The converse is obviously not true.

Criterion Selection

The term "criterion" is often used in measurement terminology to denote a predicted variable, particularly in discussions relating to the question of classical test validity. In this discussion, however, a criterion means either a cutting score or a limiting value of a proficiency range. For example, on a five item test one might set an error criterion of 2 so that pupils who have 0 or 1 errors are classified into one instructional group while those having 2 or more errors are classified into another instructional group. A similar but formally different illustration involves hypothesis testing. Suppose one wishes to classify learners into a high or low proficiency category. The extreme limits of proficiency are determined naturally: those who always get every problem right are obviously masters and those who always get every problem wrong are nonmasters. But it is also reasonable to allow for some variation in behavior so that the mastery range might extend from perfect performance, p_0, (zero error rate) up to some value p_M, for example, which denotes the maximum proportion of errors allowable in the range of performance definitely considered as mastery. The value p_M serves as an upper bound to this proficiency range. Similarly, nonmastery may be defined to include the range of proficiency from 100% error rate, p_1, down to some value p_N, the least error rate definitely considered as an indication of nonmastery. The values p_M and p_N are criterion values used in hypothesis testing associated with the "Quality-control Function."

This raises the question of how one selects criterion values. A survey of existing systems indicates a tendency to specify a rigid criterion

selection policy. Usually criteria are indicated by stating percent values such as 80%, 90%, or 100% as the minimum acceptable level of mastery performance. Analysis of sampling plans is rarely performed and one often finds little attention given to the decision-implications inherent in the casual selection of test length together with a fixed-criterion policy. It is not difficult to find instances where higher criteria are selected in the mistaken belief that this will result in a better quality of learning product than will a system having a lower criterion.

Briefly, one can use the Bernoulli model to set limits of acceptable error in classifying students as masters or nonmasters. A student with least mastery proficiency, say ζ_1, stands in greatest danger of falling below the performance criterion. Thus if $\underline{c}$ errors define the maximum allowable number of errors for masters, and $\underline{w}$ is the observed number of errors, then

$$\alpha = \sum_{w=c}^{n} \binom{n}{w} \zeta_1^{n-w} (1-\zeta_1)^{w}$$

is the probability of the CRT providing a false negative result.

Similarly, the probability of a false positive result is given by

$$\beta = \sum_{w=0}^{c-1} \binom{n}{w} \zeta_2^{n-w} (1-\zeta_2)^{w}$$

where ζ_1 and ζ_2 are arbitrary but predefined bounds to the range of mastery and nonmastery proficiency respectively.

For a student of any proficiency, ζ, the probability that his score will meet the error criterion $\underline{c}$ is

$$S = \text{Prob}\ (w<c) = \sum_{w=0}^{c-1} \binom{n}{w} \zeta^{n-w} (1-\zeta)^{w}$$

A graph of S vs. ζ clearly depends on the parameters n and c, the number of items on the test and the error criterion. Examples for various values of n and c are shown in Figures 1 - 3.

I want to make two final observations regarding this sketch of the theory underlying CRT design.

1. The more items one uses, the closer the test's classification characteristic approaches a step function. Thus, if the difference in proficiency limits between master and nonmaster is small as, e.g., proficiency in basic facts, one needs tests of approximately 20 to 25 items to separate the groups given predetermined values of α and β. Tests of greater length are likely to be wasteful of time and energy for most practical instructional decision situations.

2. The error criterion c sets the point on the proficiency scale where the characteristic curve declines most rapidly. Converted to a percentage, c should therefore fall about midway between the limiting proficiency for masters and nonmasters.

To set 100% as the criterion is equivalent to setting the error criterion c = 1. In all cases this leads to a very high probability α of false negatives. To set any fixed percentage, say 20%, as error criterion irrespective of test length is not uniformly desirable since it implicitly changes α and β for tests of different length n. The better procedure, from a theoretical point of view, is to select n and c that give approximately desirable values of α and β at the points on the proficiency scale which mark the boundaries of expected performance for masters and nonmasters.

IV. SUMMARY

In summary, the measurement information generated by CRT's is designed for use in instructional management systems where classifications of pupils for treatment are to be decided on the basis of minimal data consistent with predetermined limits for the errors of misclassification. The measures obtained are content-specific estimates of proficiency useful for the stratification of learning groups on a day-to-day basis if need be. By sampling across items rather than across persons, absolute measures of proficiency are obtained which can be reliably interpreted for nonrandomly selected pupils, the pupils of particular instructional concern. The model is designed for wide variety of applications but retains in the concept of proficiency a simple and useful index for instructional management. The empirical data generated have clear implications for instructional decision-making.

Some of the applications of item-sampling theory include the ability to (1) categorize learners into temporary learning groups on the basis of a common requirement for instructional treatment (Diagnosis and Prescription Function); (2) assess the relative effectiveness of competing instructional treatments (Instructional Assessment Function); (3) to determine, in the case of established instructional segments having predetermined performance standards, which individuals have acquired minimal standards of proficiency required for mastery and which learners require further prescriptive assistance (Quality Control Function); and (4) in the case of curriculum development, to indicate hierarchical relations within a content sequence (Curriculum Design Function).

REFERENCES

Cornfield, J. and Tukey, J., Average values of mean squares in factorials. Annals of Mathematical Statistics, 1956, 27.

Ebel, R. L., Content standard scores. Educational and Psychological Measurement, 1962, 22, 15-25.

Glaser, R., Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 1963, 18, 519-521.

Hively, W. II, et al, Generalizability of performance by job corp trainees on a universe-defined system of achievement tests in elementary mathematics calculation. Paper presented at the annual meeting of the American Educational Research Association, Chicago, February 1968.

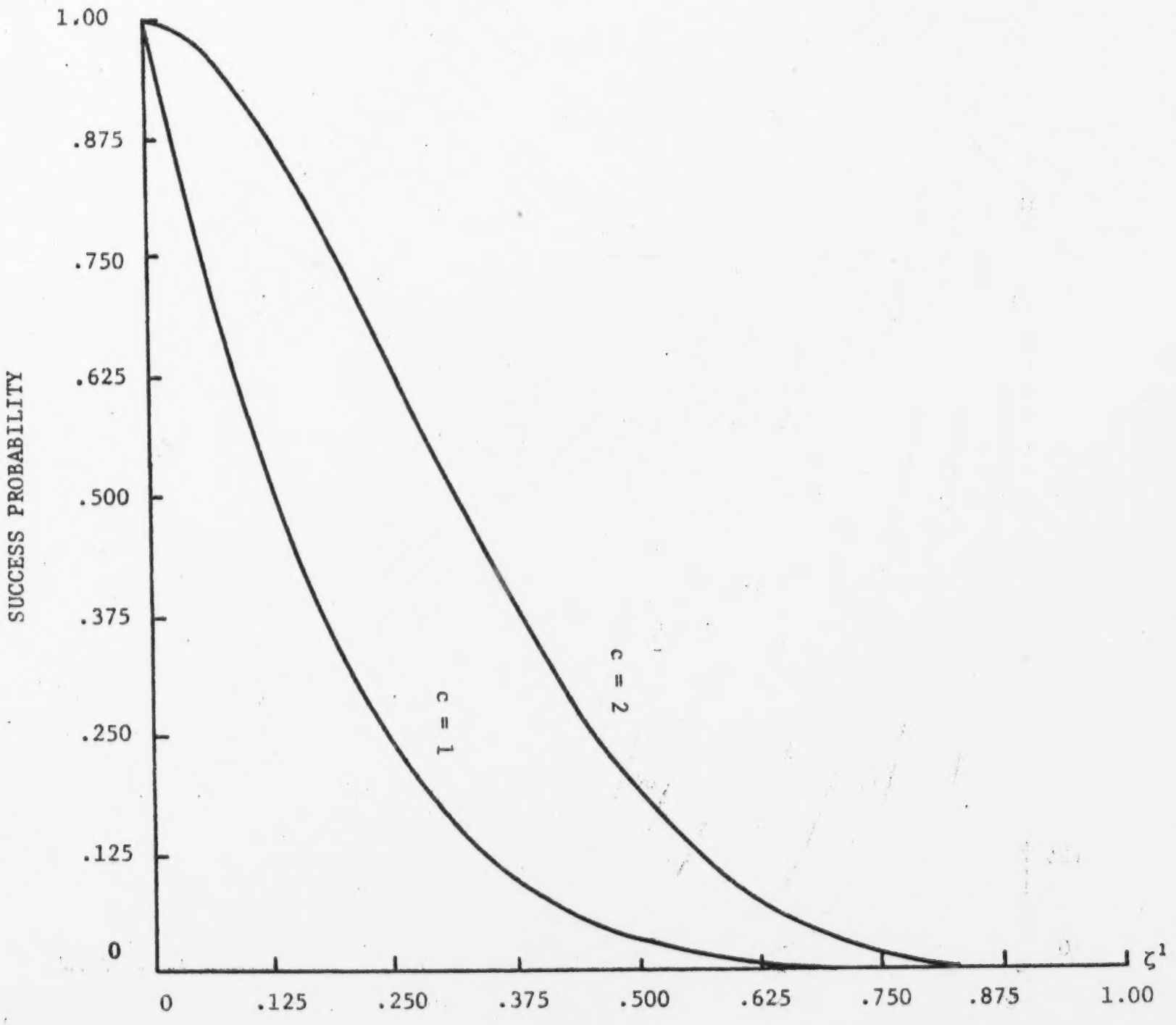
Kriewall, T, Applications of information theory and acceptance sampling principles to the management of mathematics instruction. Technical Report 103. The Wisconsin Research and Development Center for Cognitive Learning, Madison, Wisconsin, October 1969.

Lord, F. and Novick, M., Statistical theories of mental test scores. Reading, Massachusetts, Addison-Wesley, 1968.

MacDonald, J. B., Myths about instruction. Educational Leadership, May 1965.

Walbesser, H. and Carter, H., Differences in group and individually administered tests of the same behavior. Paper presented at AERA meeting, Los Angeles, February 1969.

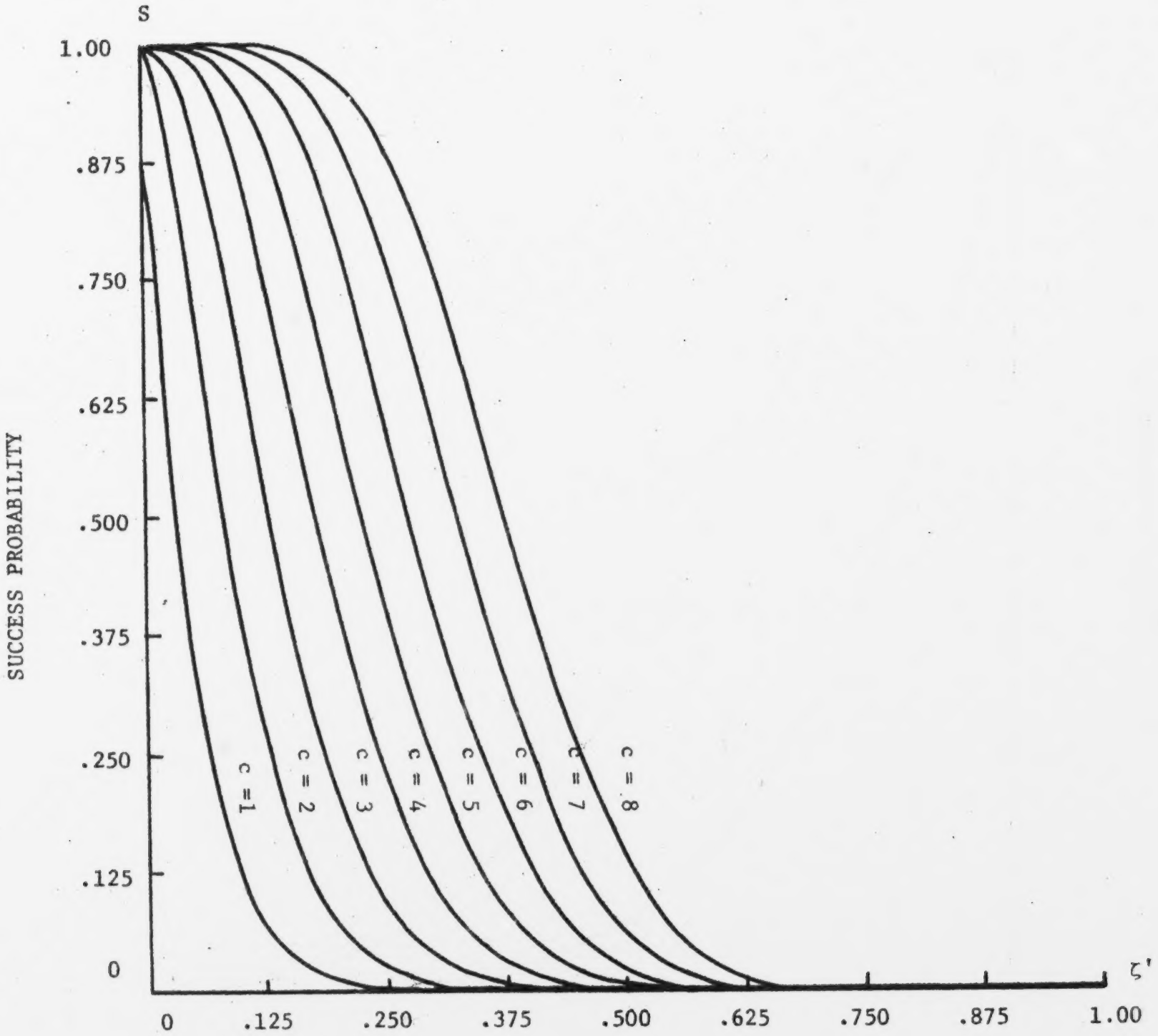
Wald, A. Sequential analysis. New York: Wiley, 1947.



n = 5

OPERATING CHARACTERISTIC (OC) CURVES FOR
FIXED N AND SELECTED CRITERIA

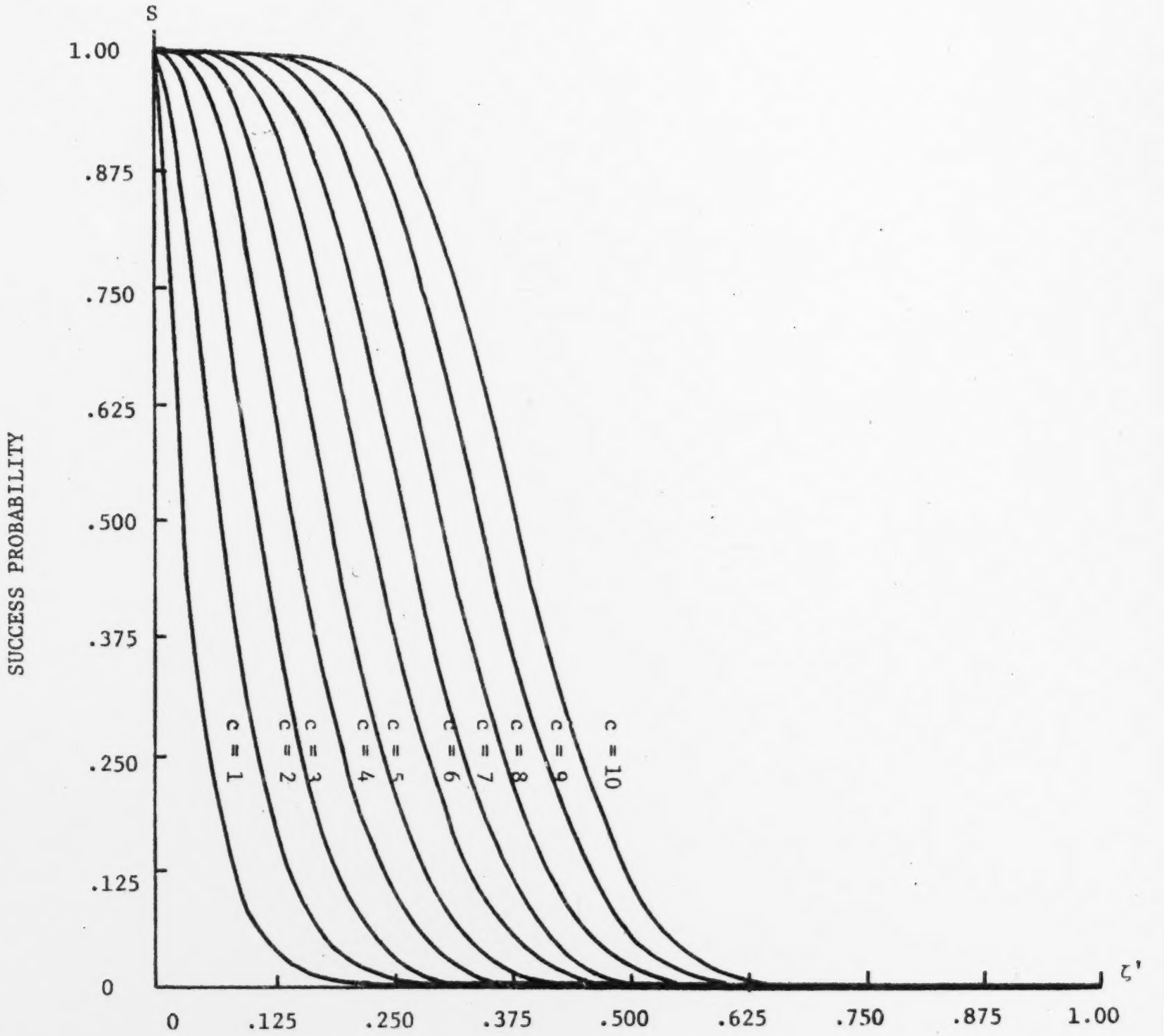
Figure 1



n = 20

OC CURVES FOR FIXED N AND SELECTED CRITERIA

Figure 2



n = 25

OC CURVES FOR FIXED N AND SELECTED CRITERIA

Figure 3